知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

科学文献与科学数据的融合方法与实例研究

◎ 姜恩波[1,2] 裴玉香[3]

[1] 中国科学院成都文献情报中心 成都 610041
[2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190
[3] 烟台大学图书馆 烟台 264005

摘要：[目的 / 意义] 关注开放科学运动的发展以及科学数据对科学研究的影响，介绍科学文献与科学数据融合的实例，阐述科学文献与科学数据融合的方法与困难。[方法 / 过程] 描述文献与数据分离的现状，说明二者融合的推动因素，通过案例介绍了科学文献与科学数据融合的 3 种呈现方式。[结果 / 结论] 科学文献与科学数据的融合是科学研究本身的一种需要，同时也是开放科学与大数据对现代科学研究的一种影响形式。在实际应用中主要有“硬关联” “软关联”和“深度融合”三种方式。科学文献与科学数据的深度融合需要顶级学术机构的综合性措施来推动。

关键词：科学文献 科学数据 科学交流 学术出版 开放科学运动

分类号：G20

引用格式：姜恩波 , 裴玉香 . 科学文献与科学数据的融合方法与实例研究 [J/OL]. 知识管理论坛 , 2019, 4(2): 69-79[引用日期]. http://www.kmf.ac.cn/p/164/.

1 科学文献与科学数据的关系

科学数据主要包括在自然科学、工程技术科学等领域，通过基础研究、应用研究、试验开发等活动产生的数据，以及通过观测监测、考察调查、检验检测等方式取得并用于科学研究活动的原始数据及其衍生数据[1]。科学的发展历程表明科学研究结果和科学研究过程同等重要。如果把一个特定研究阶段的文献产出与发表看作是科研结果的话，那么科学数据则既可以看作是科研最终成果，也可以看作是科研过程产物。论文以总结的角度对事务的背景、目的、过程和结果进行介绍和分析；而数据则详实地记录了科研的每一个过程和结果，形成客观的数据世界（digital space），是文献内容的佐证。

当前，科学数据的表现形式也在不断变化，从数据产生的渠道来看，包括但不限于观

基金项目：本文系国家重点研发计划重点专项项目“专业内容知识服务众智平台与应用示范”（项目编号：2017YFB1402400）和中国科学院文献情报能力建设专项“开放知识资源中心体系建设”研究成果之一。

作者简介：姜恩波 (ORCID：0000-0001-7890-9917)，研究馆员，E-mail：jiangeb@clas.ac.cn；裴玉香（ORCID：0000-0002-4496-4608），副研究馆员。

收稿日期：2018-07-26 发表日期：2019-04-08 本文责任编辑：刘远颖



测数据，如气象数据、天文数据、海洋生态数据，以及电子病历、穿戴设备采集数据等单体体量小但总体体量极大的个人数据；实验数据，如药物数据、基因数据、蛋白质相互作用数据等；另外，还有在这些一手数据的基础上产生的分析数据、统计数据、图表数据以及图片、音频、视频数据。多种类型的数据蕴含了极为丰富的信息。

17、18 世纪，科学期刊的出现和发展，为当时科学知识的涌现做出了突出的贡献。然而，在相当长的时间里，由于认识和技术等多种原因，科学文献和科学数据并未能够很好地关联在一起。科学文献以其较好的可阅读性、可获得性、可传播性在整个学术交流体系中占据了重要的位置，而科学数据则仍然“藏在深闺人不知”。科学实践告诉我们，在“数据密集型”科研模式之下，科学文献已经无法单独满足科研人员对科学研究的需要。科学的发展和创新需要科学理论与科学实证的融合。2012 年，英国皇家学会发布了《科学：开放事业》报告，其中提到，“一篇完整的学术论文应该包括对实验的完整描述、结果数据、不确定性评价和确保数据能被验证和重复使用的元数据”[2]。欧盟在 2016、2017 年接连发布了 Horizon 2020 框架下面向 FAIR（Findable、Accessible、Interoperable、Reusable）原则的期刊论文和研究数据管理条例与规范。在这些规范中，欧盟明确了基于公共资金支持的研究成果的管理路线图 (见图 1)[3]，并且将研究数据与期刊论文等文献成果放在了同等重要的地位，“要确保期刊论文的开放获取，并且积极推动研究数据的开放管理”[4]。开放科学的目标之一就是文献与数据共存，“所有文献在线，所有数据在线，二者皆可获得、可操作”[2]。

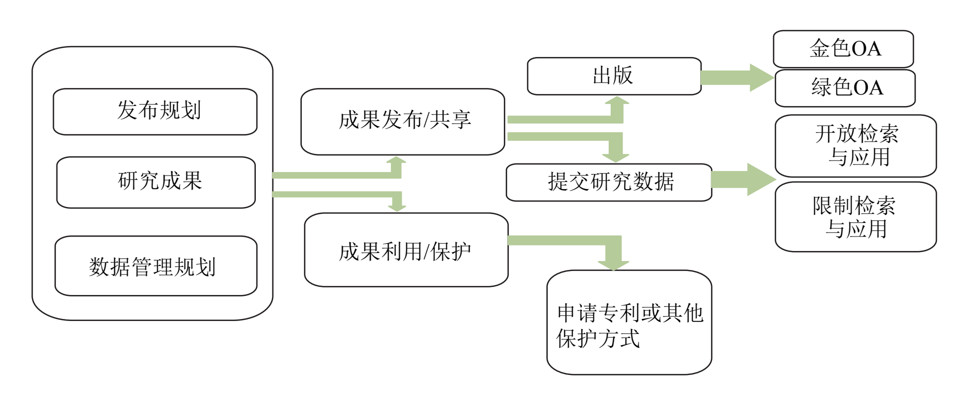


图 1 Horizon 2020 科研成果管理路线图

基于此，在当前大数据与开放运动的背景之下，科学文献与科学数据的融合和相互彰显已经成为一种趋势和需要。在后续内容中，本文将主要介绍当前出现的一些科学文献与科学数据融合的实例与方法，并在此基础上说明这种融合存在的困难。

2 大数据与开放运动下的科学文献与科学数据

2.1 大数据时代科学研究范式的改变

随着智能技术和网络技术的发展，数据规模发生了爆炸性增长。人们利用、消费着数据，同时也在产生着数据，人类进入了大数据时代。据悉，2019 年正式投入运行的 LAMOST 光谱巡天望远镜每晚产生 20 GB 的光谱数据。地面广角相机阵 GWAC 每天的观测数据量可达 7.4 TB。“天籁计划”大型射电干涉仪阵列一期 96 面天线的数据流量为 4.8 GB/s，二期 1 000 面天线的数据流量为 3.2 TB/s[5]。面对蜂拥而来且无处不在的数据，科学研究也不可避免会受到影响。2007 年，吉姆·格林 (J. Gray) 在学术报告中将科学分为经验科学、理论科学、计算模拟科

知识管理论坛，2019（2）:69-79
DOI：10.13266/j.issn.2095-5472.2019.008

学和当代的“数据密集型的科学”[6]。2009 年，美国微软研究院出版了《第四范式：数据密集型科学发现》。人们认识到数据在科学研究中的作用已经从近代科学的量化、精准和辅助决策工具发展到“不确定性”，甚至成为研究许多复杂现象主要乃至唯一的途径。

2.2 开放运动下科学文献与数据的管理政策与实践

数据与文献融合的前提是开放。只有文献和数据都能够充分地开放、方便地获取，使用者能够“以任何形式，复制、使用、分发、传递、展示原作品”[7]，二者之间的融合才会发生。鉴于此，政府、科研、出版等机构纷纷采取行动，意在推动科学文献与科学数据的开放。

2003 年，世界经济合作与发展组织（OECD）提倡所有获得公共财政资金支持的研究数据应能被公众获取。2007 年，OECD 发表《公共资助可续数据开放获取的原则和指南》[8]。同年，《柏林宣言：科学与人文科学知识的开放获取》[7] 发布，积极推动“文献”“教育”“科学数据”三类资源的开放使用，开放获取运动达到了阶段性的高峰。

政府领域，白宫科学与技术政策办公室（OSTP）、美国国立卫生研究院（NIH）、美国自然科学基金会（NSF）、欧盟、英国研究理事会等，纷纷发布研究数据共享政策。这些政策的总体理念都是要求受公共资金资助的科研项目，所形成的科学数据都应该在不妨碍国家安全、不泄露个人隐私的前提下提交、存储并提供公共访问，让用户能够免费地获取、应用以及传播。中国政府 2015 年也发布了此类政策。《促进大数据发展行动纲要》要求“加快各级政府数据开放共享，推动资源整合，提升治理能力和管理水平”[9]。2018 年 4 月 2 日，国务院办公厅发布《科学数据管理办法》，推动科学数据的汇交与共享。

学术出版领域，越来越多的学术期刊要求作者在投稿时必须向期刊编辑和同行评审专家提供相关的科学数据或者提供数据的第三方平台唯一标识符。在科学文献方面，以 PMD、BMD、ArXiv 为代表的开放获取资源在领域内的影响越来越大。以学术期刊和论文为代表的开放获取学术信息资源 (开放学术资源) 已成为学术研究不可或缺的资源，正逐步逼近“成为学术研究主流资源”的转折点 [10]。众多商业出版社也积极进入开放获取期刊出版领域。美国物理学会杂志 (The Journal of Physical and Chemical Reference Data) 从 20 世纪 70 年代早期就开始描述物理和化学材料的一般特性，目前仍在出版。2014 年 5 月，自然出版集团推出了旨在帮助科研人员发布、发现和重用研究数据的期刊《科学数据》，对研究数据的开放起到了里程碑性质的推动作用。

在互联网领域内，W3C 的关联数据云图（Linked Open Data，LOD）从 2007 年的 12 个数据集发展到 2018 年的 1 205 个数据集 (见图 2)[11]。各类原始数据通过上亿条 RDF 三元组得以发布出来，已经成为最大规模的开放数据最佳实践。另外，大量机构知识库的出现，成为开放获取实现的绿色通道。截至 2018 年 6 月，OpenDOAR 登记的机构知识库数量已经超过 3 500 个 [12]。在科学文献方面，基于 OAI-PMH、OpenURL、RSS 等协议和各类 API 的开放，众多的基于开放获取资源的集成服务平台不断涌现，如 BASE。在开放数据方面，开放数据仓储作为开放数据服务的基础，建立了规范的提交、存储和发布机制与流程，更好地将所蕴含的内容充分挖掘与利用。开放科学数据仓储的产生与发展反映了开放科学数据数量增长与人们对开放科学数据利用的需求增加。在这个基础上，出现了 re3data、Dataverse、-Datacite、Dryad 等研究数据服务平台和研究数据服务平台的登记系统。

随着各类推动科技资源开放的政策制度的颁布与运行，越来越多的科技文献与科学数据为人们所使用。科学发展内在的严谨性以及学术出版和信息服务领域的拓展，都对科学文献和科学数据的融合提出了要求。



知识管理论坛
2019 年 第 2 期(总第 20 期)

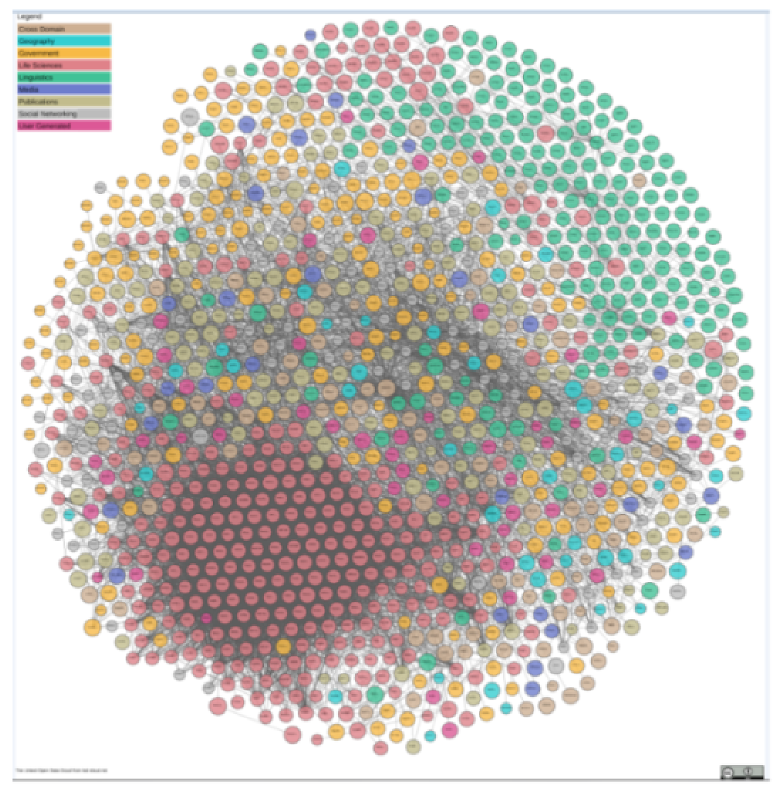


图 2 开放关联数据云图 (2018)

3 科学文献与科学数据融合类型

总体来说，当前文献与数据的融合可以归纳为 3 种类型：一种类型是基于形式的硬关联（hard connection）；一种类型是内容上的软关联（soft connection）；第三种是细粒度知识的融合 (deep integration)。

3.1 基于形式的硬关联

研究数据仓储是硬关联的代表。研究数据仓储与机构仓储类似，是用于登记、发布和存储科学数据的一种网络平台。通常来说，科研人员、期刊出版机构是研究数据仓储的两类主要用户。面向科研人员，数据仓储提供便捷的个人研究数据管理以及与文献的关联；面向期刊出版机构，数据仓储提供无缝的数据提交、管理和存储服务 [13]。研究数据仓储的特点在于开放性强、有完善的数据管理政策、充裕的存储空间和强大的技术支持等 [14]。研究数据仓储能够提供数据(集)的创建、提交、发布、引用、存储、发现和在线统计等功能。典型的研究数据仓储包括 Dryad、Figshare、R3Data。

Dryad 最初定位于生物医学领域的科学数据存储与发布。近几年业务发展较为迅速，2015 年已经与 80 种期刊达成了合作伙伴关系 [15]。2017 年，期刊合作伙伴已经达到 120 家。例如 Dryad 已经和 PLoS 全部期刊相连接，将数据存储与论文提交过程相结合。作者向 PLoS 提交论文的同时，数据会同步至 Dryad。2015 年，Dryad 发布了近 4 000 个数据包 (data package)。2017 年，Dryad 发布了其第 20 000 个数据包，并且这些数据关联到了 6 000 多种期刊 [15]。

如图 2 所示，Dryad 以数据集作为主要的描述对象。在描述数据的时候，同时也提供了文献的元数据与 URI，让使用者能够方便地跳转。Dryad 提供的是数据与文献之间一对一的关联，并且这种关系是人为建立的，通常把这种关联叫做“硬关联”。

知识管理论坛，2019（2）:69-79
DOI：10.13266/j.issn.2095-5472.2019.008

When using this data, please cite the original publication:

Jiang Y, Wang T, Wu Y, Hu R, Huang K, Shao X (2018) Past distribution of epiphyllous liverworts in China: The usability of historical data. Ecology and Evolution, online in advance of print. https://doi.org/10.1002/ece3.4274

Additionally, please cite the Dryad data package:

Jiang Y, Wang T, Wu Y, Hu R, Huang K, Shao X (2018) Data from: Past distribution of epiphyllous liverworts in China: the usability of historical data. Dryad Digital Repository. https://doi.org/10.5061/dryad.2jr1648

Cite | Share

图 3 Dryad 数据集描述信息



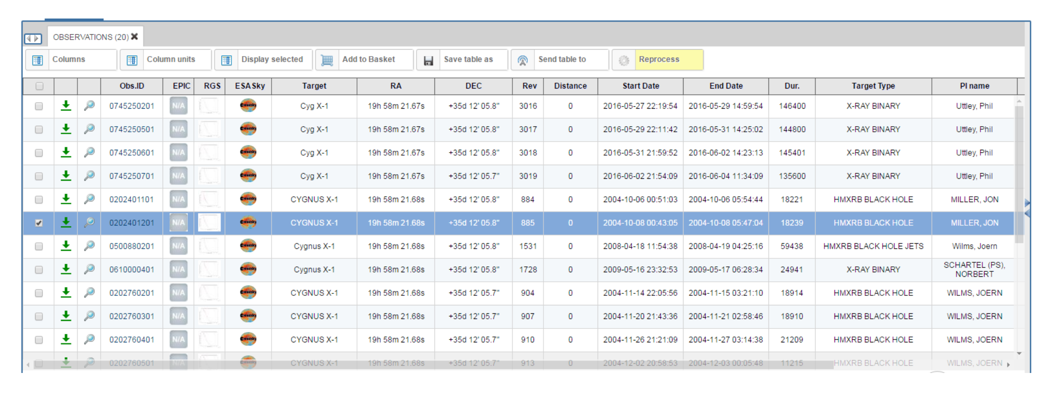
图 4 与 Dryad 数据集关联的论文信息

3.2 基于内容的软关联

基于内容的关联是当前科学文献和科学数据最为普遍的一种关联形式。以天文学为例，天文学涉及天体动力、天体物理、天体测量等，是典型的科研仪器密集型、科学数据密集型领域。难能可贵的是，天文领域在数据的开放共享、大数据管理以及综合科研信息化环境建设方面一直都很积极和规范。天文学领域的诸多数据库， 如 ADS、CDS、Chandra X-ray Center、XMM-Newton Science Archive，都把观测数据与科学文件作了很好的映射。下面以 XMM-Newton 以及 CDS 为例进行说明。

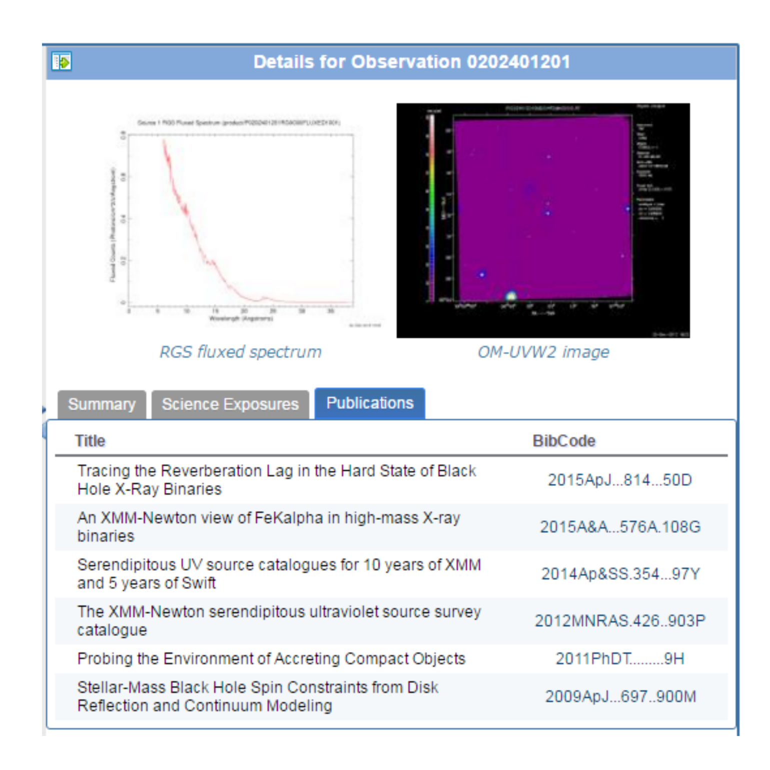
XMM-Newton 是欧洲航天局（ESA）X 射线多镜任务的建设成果，始于 1999 年。现在 XMM-Newton 是 ESA 地平线计划的科学基石之一[16]。如图 5、6 所示，在 XMM-Newton 的搜索界面输入检索词“Cygnus X-1”，检索结果包含若干行观测数据。点击其中一条查看详细信息，就能看到和这条数据图谱、色谱以及相关的出版物。这些出版物的出版年代跨度从 2009 年到 2015 年。点击年代信息，会自动跳转到 CDS 数据库获取电子版全文。而 CDS 又提供了这篇文献所包括的 10 种 SIMBAD 对象信息以及从 1850 年到现在涉及到这些对象的文献信息[17]。

从 XMM-Newton 和 ADS 两个数据库的信息来看，数据和文献的对应关系是一对多、多对一的关系。这种关联的建立并非是数据出自某一篇文献，而是这些文献中都论述到了这一方面的信息。笔者把这种类型的关联叫做“软关联”。这种软关联能够提供给用户较多的参考文献，但同时也存在不十分准确的可能性。



Obs.ID	EPIC	RGS	ESASky	Target	RA	DEC	Rev	Distance	Start Date	End Date	Dur.	Target Type	PI name
0745250201	N/A			Cyg X-1	19h 58m 21.67s	+35d 12' 05.8"	3016	0	2016-05-27 22:19:54	2016-05-29 14:59:54	146400	X-RAY BINARY	Uttley, Phil
0745250501	N/A			Cyg X-1	19h 58m 21.67s	+35d 12' 05.8"	3017	0	2016-05-29 22:11:42	2016-05-31 14:25:02	144800	X-RAY BINARY	Uttley, Phil
0745250601	N/A			Cyg X-1	19h 58m 21.67s	+35d 12' 05.8"	3018	0	2016-05-31 21:59:52	2016-06-02 14:23:13	145401	X-RAY BINARY	Uttley, Phil
0745250701	N/A			Cyg X-1	19h 58m 21.67s	+35d 12' 05.7"	3019	0	2016-06-02 21:54:09	2016-06-04 11:34:09	135600	X-RAY BINARY	Uttley, Phil
0202401101	N/A			CYGNUS X-1	19h 58m 21.68s	+35d 12' 05.8"	884	0	2004-10-06 00:51:03	2004-10-06 05:54:44	18221	HMXRB BLACK HOLE	MILLER, JON
0202401201	N/A			CYGNUS X-1	19h 58m 21.68s	+35d 12' 05.8"	885	0	2004-10-08 00:43:05	2004-10-08 05:47:04	18239	HMXRB BLACK HOLE	MILLER, JON
0500880201	N/A			Cygnus X-1	19h 58m 21.68s	+35d 12' 05.8"	1531	0	2008-04-18 11:54:38	2008-04-19 04:25:16	59438	HMXRB BLACK HOLE JETS	Wilms, Joern
0610000401	N/A			Cygnus X-1	19h 58m 21.68s	+35d 12' 05.8"	1728	0	2009-05-16 23:32:53	2009-05-17 06:28:34	24941	X-RAY BINARY	SCHARTEL (PS), NORBERT
0202760201	N/A			CYGNUS X-1	19h 58m 21.68s	+35d 12' 05.7"	904	0	2004-11-14 22:05:56	2004-11-15 03:21:10	18914	HMXRB BLACK HOLE	WILMS, JOERN
0202760301	N/A			CYGNUS X-1	19h 58m 21.68s	+35d 12' 05.7"	907	0	2004-11-20 21:43:36	2004-11-21 02:58:46	18910	HMXRB BLACK HOLE	WILMS, JOERN
0202760401	N/A			CYGNUS X-1	19h 58m 21.68s	+35d 12' 05.7"	910	0	2004-11-26 21:21:09	2004-11-27 03:14:38	21209	HMXRB BLACK HOLE	WILMS, JOERN
0202760501	N/A			CYGNUS X-1	19h 58m 21.68s	+35d 12' 05.7"	913	0	2004-12-02 20:58:53	2004-12-03 00:05:48	11215	HMXRB BLACK HOLE	WILMS, JOERN

图 5 XMM–Newton 观测数据信息



Title	BibCode
Tracing the Reverberation Lag in the Hard State of Black Hole X-Ray Binaries	2015ApJ...814...50D
An XMM-Newton view of FeKalpha in high-mass X-ray binaries	2015A&A...576A.108G
Serendipitous UV source catalogues for 10 years of XMM and 5 years of Swift	2014Ap&SS.354...97Y
The XMM-Newton serendipitous ultraviolet source survey catalogue	2012MNRAS.426..903P
Probing the Environment of Accreting Compact Objects	2011PhDT.........9H
Stellar-Mass Black Hole Spin Constraints from Disk Reflection and Continuum Modeling	2009ApJ...697..900M

图 6 XMM–Newton 观测数据出版物信息

3.3 基于语义出版的融合

如果说搜索引擎是第一代互联网的核心技术，那么语义技术则是第二代互联网的核心。2009 年，戴维·肖顿（David Shotton）提出了“语义出版”（Semantic Publishing）的概念。他认为语义出版是学术出版的一种增强。它利用互联网技术与标准（例如知识组织体系、本体、RDF、可视化、唯一标识符）来增强内容之间的互操作性，让出版的内容更加丰富和有深度 [18]。

当前，大型出版商逐步开始了语义网应用实验与服务。例如爱思唯尔（Elsevier）的“学术论文的未来” (Article of the Future) 项目、美国公共科学图书馆（PLOS）的 Semantic Enriching 项目、英国皇家化学学会（RSC）的 Prospect 项目。

如图 7 所示，爱思唯尔的“Article of the Future”语义出版理念包含以下 3 个方面 [19]：即

知识管理论坛，2019 (2) :69–79
DOI：10.13266/j.issn.2095–5472.2019.008

①呈现形式方面，提供最佳在线浏览及阅读体验；②整合内容方面，作者可以分享的更多，比如数据、代码、多媒体信息等；③相关信息方面，在线文章与来源可靠的科技信息链接，并在相关信息中呈现出来，提升附加值。如图 8 所示 [20]，整合页面由传统的“中—右”两栏变成了“左—中—右”三栏式。除了展示题录信息以外，页面还包括了章节信息、附图信息、表格信息、相关文献、引用文献甚至还有替代计量学的数据。摘要部分除了文字摘要外，还有基于图的摘要 (图 9[21])、，甚至是一段 Youtube 的视频 (图 10[22])。一篇文章在形式和内容上都被分解成为细粒度的知识单元，并且相互关联形成网络。

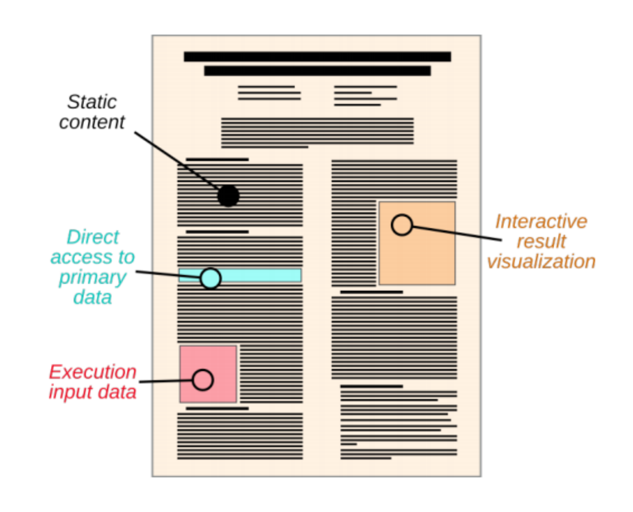
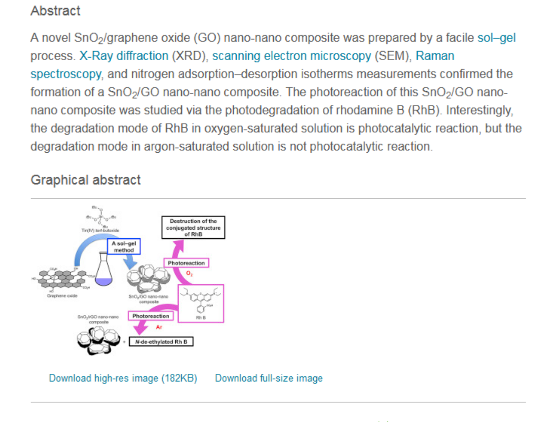


图 7 爱思唯尔 Article of the Future



图 8 Science Direct 论文展示界面


Abstract

A novel SnO_2/graphene oxide (GO) nano-nano composite was prepared by a facile sol–gel process. X-Ray diffraction (XRD), scanning electron microscopy (SEM), Raman spectroscopy, and nitrogen adsorption–desorption isotherms measurements confirmed the formation of a SnO_2/GO nano-nano composite. The photoreaction of this SnO_2/GO nano-nano composite was studied via the photodegradation of rhodamine B (RhB). Interestingly, the degradation mode of RhB in oxygen-saturated solution is photocatalytic reaction, but the degradation mode in argon-saturated solution is not photocatalytic reaction.

Graphical abstract

图 9 Science Direct 图摘要

再以英国皇家化学会（RSC）出版平台为例，来说明语义出版是如何将科学文献与科学数据相关联的。作为化学领域的资深出版商，RSC 积累了丰富的文献、数据、术语、结构式、图片等。基于此，RSC 在语义出版方面提出了“生动的科学”(Science Come Alive) 的概念。

如图 11 所示，在论文正文中有一些单词以高亮显示，并且添加了超链接。以 HPEPS 为例，HPEPS 学名为 4- 羟乙基哌嗪乙磺酸，它是一种氢离子缓冲剂。如果阅读到此处想要了解这种化学物质的基本特性、物理、化学参数，传

统阅读方式需要跳出当前阅读环境，通过相关的工具或网站去查询。而 Rich HTML 则不同，只要用鼠标点击图中的蓝色文字，在右侧则会出现有关 HPEPS 的二维、三维分子结构图以及分子式、分子量、CAS 号等相关参数信息。这样，读者的阅读过程既有纵向，也有横向，但是思维始终保持着连贯，并且在一个页面里就能阅读到通常需要到多个地方才能获取的信息。

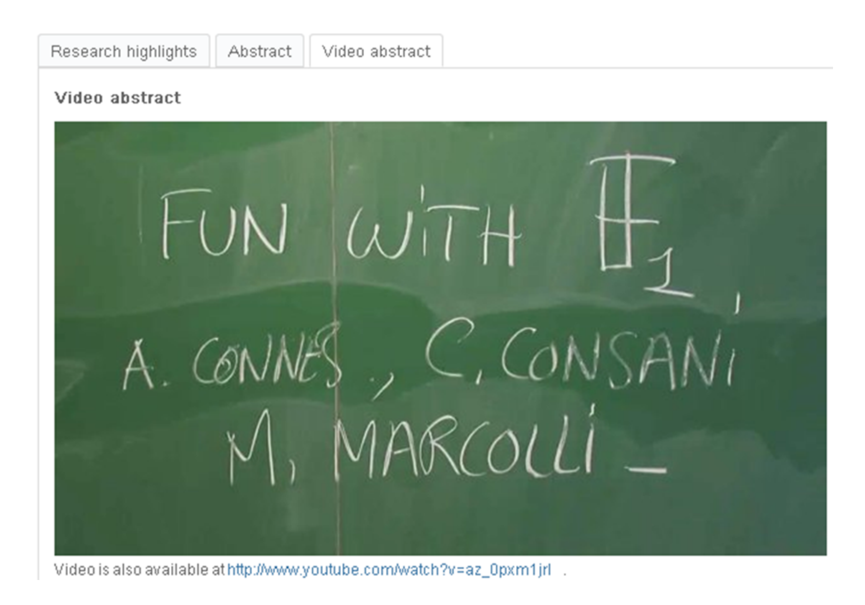


图 10 Science Direct 视频摘要

Metabolic substrate consumption and non-gaseous fermentation product formation were followed applying high performance liquid chromatography (HPLC) analysis. The HPLC (Knauer, Berlin, Germany) was equipped with a Rezex™ ROA-Organic Acid column in combination with the SecurityGuard™ cartridge AJO-4490 (Phenomenex®, Aschaffenburg, Germany). The chromatograms were recorded at room temperature with 0.005 N sulphuric acid as the eluent; the detector was a differential refractometer.

2.6. Phylogenetic analysis of the mixed culture biofilms

The qualitative identification and characterization of the microbial biofilms was performed by Nadicom GmbH Microbiology Services, Marburg, Germany, based on polymerase chain reaction (PCR)-based methods on DNA extracted from the biofilm samples. The phylogenetic analysis of the wastewater inoculum based biofilms revealed 14 genealogical trees with in a whole about 500 bacterial species (data not shown).

2.7. Scanning electron microscopy

For the electron microscopy the biofilms were prepared as follows: after a fixation step (1 h in 1% glutaraldehyde, 2% paraformaldehyde, 0.2% picric acid, 10 mM HEPES (pH 7.4), and 50 mM NaN_3), the samples were treated with 2% tannic acid for 1 h, 1% osmium tetroxide for 2 h, 1% thiocarbohydrazide for 30 min, 1% osmium tetroxide over night, and with 2% uranyl acetate for 2 h with washing steps in between. The samples were dehydrated in a graded series of aqueous ethanol solutions (10–100%) and then critical point-dried via amylacetate and CO_2. Finally, samples were mounted on aluminium stubs, sputtered with gold and examined in a DSM 940A (Zeiss, Oberkochen, Germany).

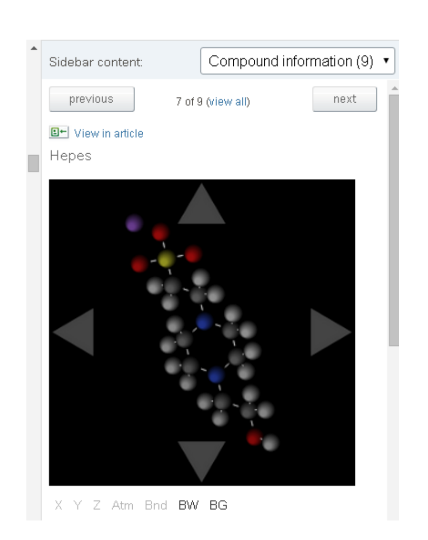


图 11 RSC Rich HTML 样例

由此可见，语义出版的实质就是知识网络。网络的最底层是本体、术语、词表，简而言之，是一个庞大的学术环境。而网络的上层则是论文中的知识点。语义出版对文献内容进行基于知识单元的抽取、规范，并把抽取出来的条目与底层的学术环境进行关联，最后通过代码和特效将这些成果都呈现在用户的阅读环境中。语义出版使得出版商提供的不再仅仅是文献的出版服务，而是一种全新的、生动的学术交流与知识传播过程。

知识管理论坛，2019（2）:69-79
DOI：10.13266/j.issn.2095-5472.2019.008

4 融合的困难与方法

科学文献与科学数据通过不同的途径产生，各自的元数据规范存在较大差异，并且在不同的语境中存在的方式也不同。科技文献和科技文献的元数据虽然是人们所熟知的，但是单就科技文献来看也存在多个种类：期刊论文、会议论文、科技报告、学位论文等。而科学数据对于图书情报以及学术出版行业来说则更为陌生。目前国内外已有一些学者对科学数据的元数据进行介绍。雪城大学秦健教授总结出科学元数据具有 4 个功能，即数据管理、数据质量控制、数据再利用和数据发现，其中数据管理功能是其他功能的基础”[23]。而不同学科领域的科学数据描述信息则更为独特，并且有该领域固有的元数据规范。例如地理学科元数据标准主要分为 FGDC 元数据标准和 ISO/TC211 元数据标准；生物多样性领域的达尔文核心（Darwin Core）元数据标准，气象领域元数据标准 CF（Climate Forecast）等。科学文献中的内容如何与这些科学数据描述信息进行匹配是融合的一个难点。因为每个领域的科学数据描述信息揭示更为充分和细粒度。而目前对科学文献内容的揭示粒度还是比较粗的。因此，除了一对一对应的“硬关联”外，要进行二者准确的关联融合就还需要对内容进行进一步的处理。

在关联的过程中，被关联的一方，即 Target 的描述更为重要。因为 Target 需要从自己的描述中找到和关联方即 Source 传来的内容相一致的内容。但从实际应用来看，从文献到数据和从数据到文献的情形都存在。因此，从数据这一部分来看，需要形成科学数据的元数据，对其内容进行完整的描述。例如数据题名、数据所包含的标准写法（例如 HPEPS、4- 羟乙基哌嗪乙磺酸）、别名、分子式、观察地点（经纬度）等。总之，需要让可关联的点更加丰富一些。而对于科学文献这一部分来看，需要从文献中抽取出特定的知识单元。毕竟数据传过来的关联需求只能是内容片段，不会是一个完整的文献名称。这些抽取出来的内容会比作者自己形成的关键词要更为丰富。

再者，对科学文献进行细粒度知识单元的抽取时，需要借助比较完备的知识组织体系，例如领域词典、叙词表、本体或者自己建立的实体（知识）库。一来是将其作为抽取的凭据，再者可以作为已有内容的规范依据。总之，这些基础都能让内容的抽取更为准确，从而提高关联和融合的准确性。

从前面的案例可以看出，目前能够建立较大规模、揭示程度较深的融合服务的机构都是站在特定行业或者学科领域顶端的机构。首先，这些机构有其独立、稳定的信息汇聚渠道。例如出版商控制学术期刊；而有的行业协会则拥有行政管理的职责，甚至自己出版学术期刊。其次，经过多年的积累，这些机构有着深厚的学术背景和技术能力。他们了解学术发展的动态，了解科研人员的需要，同时也有能力在海量资源的基础上进行更加深入的处理，从而推出更好的服务。以英国皇家学会（RSC）为例，RSC 利用自建或开放的本体（RXNO、C-MO、MOP、GO、SO 等）、化学结构数据库（ChemSpider），对文章进行细粒度标引，抽取出其中的专业概念，如化合物名称、分子式、术语、机构等，从而使 RSC 在线出版的文章极大地增强了对学科知识的揭示和关联能力。因此，科学数据与科学文献的融合不仅仅是一个技术问题，更应该是“资源的积累 + 多种技术的综合应用 + 新型用户学术服务”的结合体。

5 结论

科学文献与科学数据的关联与融合既是当前科学研究的迫切需要，也是互联网、计算机、大数据、智能设备等各项技术发展的产物，同时也是开放科学运动的阶段性成果之一。科学数据与科学文献融合，能够将研究数据在科学事业中的作用进一步发挥出来，让科学研究的链条不再缺失，同时也能够深度挖掘文献与数据的关系，为用户呈现出更为丰富的信息服务。

当前文献与数据的融合是大数据时代科研与服务的一个亮点，同时也仅仅是一个开始，期待在这个领域有更多的精彩案例与服务应用。

参考文献：

[1] 国务院办公厅关于印发科学数据管理办法的通知 [EB/OL]. [2018-09-28]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.

[2] 英国皇家学会. 科学：开放的事业 [M]. 上海：上海交通大学出版社，2015.

[3] Open/FAIR research data in Horizon 2020 [EB/OL]. [2018-12-18]. https://ec.europa.eu/easme/sites/easme-site/files/open_fair_research_data_in_h2020.pdf.

[4] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 [EB/OL]. [2018-12-18]. https://www.openaire.eu/guidelines-on-open-access-to-scientific-publications-and-research-data-in-horizon-2020.

[5] 崔辰州 . 大数据时代的天文学研究 [J]. 科学通报，2015，60(5/6): 445-449.

[6] HEY T, TANSLEY S, TOLLE K. 第四范式：数据密集型科学发现 [M]. 潘教峰，张晓林等译 . 北京：科学出版社 , 2012.

[7] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities[EB/OL]. [2018-06-25]. https://openaccess.mpg.de/Berlin-Declaration.

[8] OECD Principles and Guidelines for Access to Re- search Data from Public Funding[EB/OL]. [2018-10-10]. http://www.oecd.org/sti/sci-tech/38500813.pdf.

[9] 促进大数据发展行动纲要 [EB/OL]. [2018-05-25]. http://www.gov.cn/gongbao/content/2015/content_2929345.htm.

[10] 张晓林 , 李麟 , 刘细文 , 等 . 开放获取学术信息资源：逼近“主流化”转折点 [J]. 图书情报工作 , 2012, 56(9): 42-47.

[11] The Linked Open Data Cloud[EB/OL]. [2018-06-10]. http://lod-cloud.net/.

[12] OpenDOAR Statistics[EB/OL]. [2018-06-10]. http://v2.sherpa.ac.uk/view/repository_visualisations/1.html.

[13] The Dataverse Project[EB/OL]. [2018-06-10]. https://dataverse.org/.

[14] 殷沈琴 , 张计龙 , 张莹 , 等 . 社会科学数据管理服务平台系统选型研究 [J]. 图书情报工作 , 2013, 57(19): 92-96.

[15] Dryad Executive Director to pursue new opportunities[EB/OL]. [2018-06-12]. https://blog.datadryad.org/2018/02/.

[16] Welcome to the XMM-Newton Science Operations Centre[EB/OL]. [2018-04-26]. https://www.cosmos.esa.int/web/xmm-newton/home.

[17] Tracing the reverberation lag in the hard state of black hole X-ray binaries[EB/OL]. [2018-04-26]. http://simbad.harvard.edu/simbad/sim-ref?querymethod=bib&simbo=on&submit=submit+bibcode&bibcode=2015ApJ...814...50D.

[18] Semantic Publishing[EB/OL]. [2018-04-26]. https://semanticpublishing.wordpress.com/.

[19] The Article of the Future[EB/OL]. [2018-05-15]. https://www.elsevier.com/connect/the-article-of-the-future.

[20] MULVENNAA1 J, MOERTEL L, JONES M K, et al. Exposed proteins of the Schistosoma japonicum tegument [J]. International journal for parasitology, 2010, 40(5): 543-554.

[21] TOMOYUKI TAJIMA T, GOTO H, NISHI M, et al. A facile synthesis of a SnO_2/Graphene oxide nano-nano composite and its photoreactivity [J]. Materials chemistry and physics, 2018, 212:149-154.

[22] CONNES A, CONSANI C, MARCOLLI M. Fun with F1 [J]. Journal of number theory, 2009, 129(6): 1532-1561.

[23] 赵华 , 王健 . 国内外科学数据元数据标准及内容分析 [J]. 情报探索 , 2015(2): 21-24.

作者贡献说明：

姜恩波：确定论文选题，拟定提纲，撰写部分论文，修改论文；

裴玉香：撰写部分论文，修改论文。

ChinaXiv合作期刊
知识管理论坛，2019（2）:69-79
DOI：10.13266/j.issn.2095-5472.2019.008

Research and Practice on the Integration of Scientific literature and Scientific Data

Jiang Enbo[1,2] Pei Yuxiang[3]
[1]Chengdu Library and Information Center, Chinese Academy of Science, Chengdu 610041
[2]Department of Library Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190
[3]Yantai University Library, Yantai 264005

Abstract: [Purpose/significance] This paper focuses on the development of the open science movement and the influence of scientific data on scientific research, introduces some cases of the integration of scientific literature and scientific data, and states the methods and problems of integration. **[Method/process]** The author described the status of separation of scientific literature and scientific data, explained the background that promotes the integration of these two things. Then, by the case study, it introduced three types performance of the integration of scientific literature and scientific data. **[Result/conclusion]** The integration of scientific literature and scientific data is needed by scientific research, as well as a form of influence on modern scientific research in the era of open science and big data. In practical application, there are mainly three ways: “hard connection”, “soft connection” and “deep integration”. The integration of literature and data need to be promoted by the comprehensive measures from the main institutions of all fields.

Keywords: scientific literature scientific data scientific communication scholar publishing open science movement